Protein family classification based on searching a database of blocks

Steven Henikoff[1,2]
Jorja G. Henikoff
[1]Howard Hughes Medical Institute
Basic Sciences Division
Fred Hutchinson Cancer Research Center
Seattle, WA 98104

[2]Corresponding author
Phone: (206) 667-4515
FAX: (206) 667-5889
e-mail: henikoff@sparky.fhcrc.org

Keywords: database searching; sequence homology; protein blocks
Running title: Protein family classification

ABSTRACT

The most highly conserved regions of proteins can be represented as "blocks" of locally aligned sequence segments. Previously, an automated system was introduced to generate a database of blocks that is searched for local similarities using a sequence query. Here we describe a method for searching this database that can also reveal significant global similarities. Local and global alignments are scored independently, so they can be used in concert to infer homology. A set of 7,082 diverse sequences not represented in the database provided queries for testing this approach. The resulting distributions of scores led to guidelines for interpretation of search data and to the classification of 289 uncatalogued sequences into known groups. Thirty-eight of these relationships appear to be new discoveries. We also show how searching a database of blocks can be used to detect repeated domains and to find distinct cross-family relationships that were missed in searches of sequence databases.

INTRODUCTION

As a result of the accelerating expansion of sequence databanks, it becomes increasingly probable that a search for similarity will succeed in detecting a relationship between any newly determined sequence and one or more known sequences. Often such relationships are important clues to gene or protein function. However, sometimes the similarity is too weak for a potentially interesting relationship to be detected above the background of chance alignments. Background increases with the growth of sequence databanks, making distant relationships even more difficult to detect with confidence.

Detection of distant relationships can be aided by the presence of multiple members of a single protein family in a database. For example, BLAST3 (Altschul and Lipman, 1990) rescans a list of local alignments that score within the "twilight zone" of search results to identify significant 3-way local relationships. Alternatively, a database in which relationships are explicitly represented can be searched (Bairoch, 1992; Smith, R. F. and Smith, 1990; Henikoff and Henikoff, 1991; Harris *et al.,* 1992; Pongor *et al.,* 1993). An example of this latter approach is a database of protein "blocks" where each block is a local multiple alignment of ungapped segments from a group of related proteins (Henikoff and Henikoff, 1991). A query sequence is searched against this database of blocks by calculating a position-specific scoring matrix (Gribskov *et al.,* 1987) representing each block and scoring every possible position in the query for all blocks in the database. Searching a database of blocks provides information on local relationships, useful for identifying sequence motifs. These searches are more specific than are searches of sequence databases because blocks represent only the most highly conserved regions of proteins, a much smaller set than the set of sequences.

Most protein families are characterized by multiple local motifs indicative of more global relationships. Popular searching programs based on pairwise sequence comparisons (*e. g.,* Smith, T. F. and Waterman, 1981; Pearson, 1990; Altschul *et al.,* 1990; Gish and States, 1993), though designed to search for local relationships, can detect global relationships as well. However, current searching methods designed to detect local motifs common to multiple sequences do not take advantage of the global information implied by multiple local motifs (Altschul and Lipman, 1990; Fuchs, 1991; Wallace and Henikoff, 1992). The value of such global information for detecting or verifying a family relationship motivates the approach described here.

Global information is present in the Blocks Database as multiple blocks and distances between them observed for the sequences in the protein family. If a query sequence belongs to a family with multiple blocks, then at least a subset of these blocks should score highly in a search and be arranged in a compatible way along the query. In the present approach, we quantify the degree to which this is the case. We demonstrate the sensitivity and selectivity of this approach by the detection of uncatalogued relationships for proteins not represented in the database of blocks. Using suspected nucleic acid dependent ATPases of current biological interest as examples, we also show that interesting cross-family relationships can be readily discerned.

METHODS

We define blocks as multiply aligned sequence segments without gaps, and groups as collections of proteins that share sequence similarity. For this work, we used groups that are listed in the PROSITE catalog (Bairoch, 1992), but a database of blocks could be made from any collection of protein groups. The PROSITE catalog includes a file (PROSITE.DAT) in which each entry contains the SWISS-PROT IDs for members of a

protein group. Although each entry also provides a manually-derived PROSITE consensus pattern, this is not used in generating blocks. PROSITE groups are represented in the Blocks Database by one or more blocks (the average is 3.7 blocks) generated by the automated PROTOMAT system (Henikoff and Henikoff, 1991). The PROSITE pattern may or may not be contained in one of the blocks for a group. Calibration of individual blocks and their concatenation into a single file results in a database that can be searched using a query sequence. The current Blocks Database contains 2,302 blocks representing 619 groups. The Blocks Searcher system consists of the successive execution of two programs. PATMAT converts each block to a position-specific scoring matrix (Henikoff *et al.,* 1990) and scores all possible alignments of the DNA or protein query sequence and the Blocks Database as previously described (Wallace and Henikoff, 1992). A rank-ordered list of the individual blocks in the database is the result. A new program, BLOCKSORT, then analyzes the result of a PATMAT search by collecting the alignments for individual blocks belonging to a group and evaluating the group as a whole. The overall strategy is outlined in Fig. 1.

So that search results can be evaluated quantitatively, two technical problems are addressed in this study, one related to the scoring of individual blocks (PATMAT) and the other to the evaluation of multiple blocks representing a group (BLOCKSORT). First, individual blocks should be scored fairly in competition with one another. We show that our empirical calibration procedure leads to block scores for shuffled sequences that approach the expected distribution of scores and do not appear to favor particular blocks. Second, multiple blocks should be evaluated by a global measure that accurately reflects the chance probability of a sequence aligning correctly with them. We show that our calculated "expectant value" (E) corresponds closely to probabilities observed in searches using both shuffled and unshuffled true negative query sequences.

Determining hits

A "hit" reported by the Blocks Searcher consists of one or more blocks from a protein group represented in the Blocks Database. The blocks in a hit must be positioned in the query sequence in a manner that is compatible with their positions in the sequences documented as belonging to the group, where compatibility is determined by order and distance apart.

Individual block alignments are sorted by PATMAT score, a measure of local similarity (see below). By default, the Blocks Searcher currently saves the best 400 alignments for analysis, except for DNA sequence queries >5000 bp for which 1000 alignments are saved. BLOCKSORT first sorts the saved PATMAT results by strand and block name, determines the minimum rank for each separate BLOCKS group, and re-sorts by minimum rank, strand and block name. It then analyzes each group that has at least one block with PATMAT score above 1000, which is the 99.5 percentile calibration point (Henikoff and Henikoff, 1991). By default the maximum number of hits reported is currently set to 10; for DNA sequences larger than 10,000 bp, this number increases by one for every 1000 bp. Defaults are determined empirically to achieve a balance of sensitivity and selectivity for sequences of typical lengths (data not shown).

For each hit, the PATMAT score, rank, and location (frame and offset) of each block alignment in the group is listed. BLOCKSORT looks up the group for each hit in the Blocks Database and prints a map of the relative locations of the blocks in the database. This includes the width and order of the blocks separated by the minimum and maximum distances for all sequences represented by the blocks. Then it checks the alignments of

the blocks in the group with the query, looking for the most compatible arrangement, and builds a query map. The BLOCKSORT output for the example diagrammed in Fig. 1 is shown in Fig. 2.

The highest ranking block in a hit is called the "anchor" block (Block A in Fig. 2) and any other blocks in the hit are called "supporting" blocks (Blocks B and D). Note that if 10 hits are reported in a search, there will be 10 anchor blocks, one per hit. While the anchor block for a true positive hit is typically the highest ranking block in the search, this need not be the case.

Starting with the anchor block, BLOCKSORT considers each other block in rank order to see if it supports the anchor block. A supporting block must align with the query sequence in the correct order and within reasonable distances of the other blocks mapped so far. The distance from neighboring blocks is considered reasonable if it is at least -1 (overlaps a neighboring block by at most 1 amino acid) and at most the sum of the maximum and minimum distances for any sequence in the group. As each supporting block is added to the query map, the distances are checked to the closest blocks already in the query map on either side. In the example, Block C is excluded from the hit because it aligns 40 residues upstream from the alignment with the higher-ranking Block D, whereas the largest allowable distance (maximum + minimum) is only 34 residues (23 + 12) in the best path. Blocks from the group included in the final query map together constitute the hit. Blocks from the group that do not fit are displayed below the query map. For each block included in a hit, the alignment of the query with the sequence from the block in the Blocks Database with which it shares the most identical residues is shown to assist in evaluating the hit.

Evaluating single block hits

Each alignment of the query sequence with a block is an ungapped local alignment between the query sequence and sequences belonging to the group represented by the block. Since blocks are of different widths and different degrees of similarity, it is necessary that blocks scores be calibrated to allow comparisons to be made between them. Calibration is achieved by dividing the raw score by a "lower calibration score", resulting in a PATMAT score (Henikoff and Henikoff, 1991). The raw score is the sum of scores for each aligned position using the position-specific scoring matrix derived from the block. The lower calibration score is the 99.5th percentile level of presumed true negative alignments obtained using the same matrix to similarly score all possible alignments of the block with the SWISS-PROT protein sequence database (Bairoch and Boeckmann, 1992). Thus, an alignment of a sequence segment with a block that obtains a PATMAT score of 1000 should be as good as an alignment at the 99.5 percentile level of true negatives when that block is used to search SWISS-PROT. For single block hits, the PATMAT score is used for evaluation of the implied local similarity. For multiple block hits, ranks rather than PATMAT scores are used. This both reduces the effect of imperfections in the calibration procedure and allows a simple intuitive model to be used in estimating the chance probability of a hit, described below.

Evaluating multiple block hits

Multiple block hits contain information about global similarity between the query sequence and members of a group. We seek a measure of global similarity in addition to the PATMAT scores for individual blocks in the hit. This measure should quantify how often the blocks reported in the hit are arranged along the query sequence in the correct

order and are separated by reasonable distances. It is very difficult to construct a realistic theoretical model for scoring multiple block hits because different protein groups include different numbers of sequences and are represented in the Blocks Database by different numbers of blocks with diverse properties. Therefore, we intuitively model the process of searching a database of blocks to compute an expectant value, E, that can be used to quantify the degree of global similarity. For example, a value of $E=10^{-3}$ would be expected to occur by chance once for every 1000 searches of the database. Whether or not these expectant values are realistic is determined empirically.

Let the query sequence be of length N amino acids. If the query is a DNA sequence, N is 3 times the number of nucleotides in the sequence since the query is translated in all 3 frames on each strand and all the block alignments in a hit must be on the same strand. Let B be the number of blocks in the Blocks Database. In the search of the query against the Blocks Database, each block in the database independently receives a PATMAT score for every possible alignment with the query, so there are N ranks for each block and NB total ranks assigned. Let the particular set of blocks from which a hit is mapped have G member blocks in the Blocks Database. In the search, the G blocks in the hit are assigned NG different ranks. The block belonging to the group assigned the minimum rank (not necessarily rank 1) is the anchor block. Let the number of supporting blocks in the hit be S, S<G, the rank of each supporting block be $rank_s$, and the allowable minimum and maximum distances from the anchor block to each supporting block as computed from the sequences in the group in the Blocks Database be min_s and max_s, s=1,S. For the example shown in Figure 1, A is the anchor block and D the first supporting block, min_s is the sum of widths for blocks B and C less one to allow for overlap (26+58-1=83) and max_s is the sum of widths for blocks B and C plus the sum of minimum and maximum distances [26+58+(10+14)+(8+41)+(12+23)=192]. Let the distance between the anchor and supporting blocks in the query be $dist_s$, s=1,S. Then we model the probability that the supporting blocks could be found by chance as:
Since each alignment is scored independently, this becomes:

Since the ranks are assigned independently of the location of the alignments, this

becomes:

We estimate the rank probability by analogy to the situation in which an object is drawn at random from a collection of objects of two types without replacement (Lindgren, 1968). Such a situation might occur for example in draw poker, where a player holding an ace (analogous to the anchor block) wishes to know the probability of drawing more aces (analogous to supporting blocks) from the deck (analogous to the database). In our situation, the objects are the NB alignments of blocks from the database and the two types are the NG alignments that belong to the hit group and the NB-NG that do not. We

compute the simple probability that the S+1 alignments from the hit group are drawn in the order observed. The task is to assign ranks to NG alignments from among the NB evaluated in the search. We estimate the rank probability for this model using the hypergeometric distribution. For the s^{th} supporting block let:

D = number of alignments ranked since the previous supporting block,
T = number of alignments left to be ranked,
R = number of alignments from the group represented by the hit left to be ranked.

So the number of alignments not from the group represented by the hit left to be ranked is T-R. Then for the s^{th} supporting block:

$P(\text{rank}=\text{rank}_s)$
= P(one of the D alignments ranked is from the hit group)
= (#ways to choose 1 hit group alignment from R)
* (#ways to choose D-1 non-hit group alignments from T-R)
÷ (#ways to choose D alignments from T):

We estimate the distance probability as the fraction of allowable positions of the supporting block in the sequence:

Since each alignment of each block is ranked independently, rank and distance probabilities are computed for each supporting block separately and the probabilities for all supporting blocks are multiplied together to obtain the expectant value (E). Because our method for obtaining E is based on probabilities, $0 \leq E \leq 1$. For single block hits, where no global similarity measure is available, E=1.

Queries for empirical tests

The SWISS-PROT 24 database was used to provide a list of all sequences that were not represented in the PROSITE 10.0 catalog from which groups were obtained to generate the current Blocks Database (Blocks 6.0). To maximize diversity among test queries, only a single sequence with the same protein name but different species name was chosen, leading to the selection of 7,082 sequences. Each of these sequences was used to query Blocks 6.0, producing 46,022 hits, nearly all of which should be true negative hits. Each sequence was also shuffled by randomly permuting individual residues, and each shuffled sequence was used to query Blocks 6.0, producing 43,783 true negative hits. For convenience, we refer to the 7,082 SWISS-PROT sequences as the "true negative sequences", and to their shuffled versions as the "shuffled sequences".

Implementation

The Blocks Searcher has been implemented as an electronic mail server (Henikoff *et al.,* 1993). Detailed instructions with illustrative examples can be obtained by sending the message "help" in the subject line to blocks@howard.fhcrc.org. The Blocks Database is updated semi-annually following each significant update of PROSITE. The PATMAT and BLOCKSORT programs are written in standard C for UNIX workstations and are available by anonymous ftp from the NCBI repository, ncbi.nlm.nih.gov, in the blocks subdirectory. Further information can be obtained by sending a request to henikoff@howard.fhcrc.org.

RESULTS

Evaluating local similarity

It is important to ascertain whether the PATMAT score used as measure of local similarity can be interpreted in terms of a reasonable model of chance. Ideally, all blocks should be equally likely to score at or above a given level in searches against the Blocks Database using a random query sequence. Since it is difficult to construct random protein sequences sufficiently similar to real sequences in length and composition, we chose to test a set of fictitious sequences with lengths and compositions identical to a diverse set of real sequences by shuffling 7,082 sequences selected from SWISS-PROT.

For all searches, the PATMAT scores for the highest ranking blocks can be used to assess the effectiveness of the calibrated PATMAT score for making direct comparisons of blocks of different composition (Henikoff and Henikoff, 1991). If all blocks were equally likely to rank first, then on the average, a block should rank first in the shuffled sequence searches about 3 times (7,082 searches ÷ 2,302 blocks = 3.02), and a Poisson distribution of frequencies should result for all 2,302 blocks. However, because shuffled sequences reflect the same variations in amino acid composition and length as for the real sequences from which they are derived, even perfect calibration might not lead to a Poisson distribution of rank 1 block frequencies. Nevertheless, the observed distribution has mean 2.98 and does not seriously deviate from a Poisson distribution (Fig. 3A). Furthermore, when the search data are divided arbitrarily into two equal parts, only 2 of the 20 blocks that appear in the tail of one set appear in the tail of the other set.

Another assessment of calibration effectiveness comes from examination of the distribution of anchor blocks for shuffled sequences. A total of 43,783 true negative hits resulted from the searches of 7,082 shuffled sequences against the Blocks Database. This means that on average at least one block from about 6 different groups achieved a PATMAT score of at least 1000 in each search. Each hit includes an anchor block with a PATMAT score reflecting the best local alignment between the query and the highest ranking block in the hit. If all blocks were equally likely to appear as the anchor block in a hit, then on average each block would be the anchor for 43,783 hits/2,302 blocks = 19 hits. However, If some blocks become anchor blocks in hits at unexpectedly high (or low) frequencies, then a multiphasic distribution should result. Fig. 3B shows that the distribution is approximately normal with mean 18.5. About 2% of the blocks fall into small peaks at either end of the distribution, suggesting imperfections in the calibration procedure. However, it is noteworthy that only one of the ten blocks lying within the high-frequency tail of the distribution of rank 1 blocks (Fig. 3A) is the same as one of those lying within the tail of the distribution of all anchor blocks (Fig. 3B). Examination of these high-frequency blocks does not reveal any common feature that could account for their better performance in the searches (data not shown). Together, these observations on the

distribution of PATMAT scores for rank 1 blocks and all anchor blocks indicate that our block calibration procedure is effective in preventing some blocks from being unduly favored in a search. PATMAT scores are reported with percentiles of the distribution of the scores for the shuffled queries to aid in evaluation of local similarity. For calculating a global similarity measure, our use of ranks rather than PATMAT scores should minimize effects of imperfections in the calibration procedure, since only the order of scores matters and not their precise distribution.

Evaluating global similarity

Hits from searches of the shuffled sequences against the Blocks Database can also be employed to evaluate the expectant value used to assess global similarity for multiple block hits (Table 1a). The most significant single hit obtained an expectant value E of 2.9×10^{-5}, very close to the observed probability of hits in this range of E (1/43,783 = 2.3×10^{-5}). Furthermore, for all intervals of E, the observed probabilities of hits are very close to the value of E itself, suggesting that our expectant value E can be used as a proxy for the probability of a multiple block hit. While it is possible that the intuitive model used to calculate E is not ideal, the empirical support described in this and the next section justifies using E to estimate the significance of hits.

Together, the anchor block scores and the expectant values from the searches with shuffled queries provide independent evidence that can be used to evaluate a hit, because the anchor block score is not used to calculate E (Fig. 4). For example, a hit with anchor block score as good as or better than 1300 (≥ the 98th percentile) and expectant value as good as or better than 10^{-3} is expected to occur at least once by chance in 7000 searches, but is not expected to occur in 1000 searches. A hit with anchor block score of 1200 (≥ the 85th percentile) and expectant value of 10^{-3} is expected to occur by chance at least once in 1000 searches, but is not expected to occur in 100 searches.

New classifications

Searches were also carried out using the same 7,082 sequences without shuffling, which provided a set of 46,022 hits (Table 1b). Evaluation of these hits must take into account the possibility that many of the presumed true negatives are actually true positives, but were not catalogued as such in PROSITE 10.0. In addition, many local similarities are known but not catalogued, for example ATP-binding domains and glycine-rich regions. Therefore, the highest scoring hits were examined manually in order to remove these known true positives from the lists of results. In all, 289 hits were removed from the top of the lists. The distributions of anchor block scores and expectant values from these pruned results lists are very similar to the distributions obtained using shuffled queries (Table 1). For example, the best true negative expectant value was $E = 1.0 \times 10^{-4}$, slightly less significant than the best obtained for the shuffled sequences ($E = 2.9 \times 10^{-5}$).

High scoring true positive hits identified in the above analysis include 38 uncatalogued relationships involving 29 different protein groups that do not appear to have been reported in the original or subsequent publications (Table 2). For example, there are two hits involving genes from yeast chromosome III that were not reported in the original study presenting the complete sequence of the chromosome (Oliver *et al.,* 1992), nor in more careful studies in which a variety of methods were employed to discover relationships (Bork *et al.,* 1992b; Bork *et al.,* 1992a): The first of these is the hit that aligns YCZ2_YEAST and the zinc alcohol dehydrogenases (BL00059, Fig. 2), with $E = 1.4 \times 10^{-6}$; a value this good is expected to occur by chance only about once in 1 million searches,

with independent evidence provided by the anchor block score (98.5th percentile). Three other new members of this large and diverse family scored even better (Table 2), yet were overlooked by the authors of the original papers. The second yeast chromosome III hit aligns YCD9_YEAST with the beta-transducins (BL00678, Fig. 5A). The anchor block score is in the 99.9th percentile, with independent evidence provided by the expectant value of 0.00013 (Table 2), a combination of the two measures that is well above any that have been observed in test searches (Fig. 4).

Several other relationships are worth noting. Homology between cholesterol oxidase (CHOD_BREST) and the other flavin-dependent oxidoreductases in the GMC group (BL00623) was not detected previously (Fig. 5B, Cavener, 1992; D. Cavener, personal communication). This finding takes on added importance considering that the 3D structure of cholesterol oxidase is known (Vrielink *et al.,* 1991), and so can be used to model members of the GMC group, whose structures are unknown. Other previously unreported similarities include that between ribonuclease I (RNI_ECOLI) and the ribonuclease T2 family (BL00530, Fig. 5C), between carboxypeptidase S (CBPS_YEAST) and the diverse family that includes carboxypeptidase G (BL00758, Fig. 5D), and transporters of the oligoamines cadaverine (CADB_ECOLI) and putrescine (POTE_ECOLI) and a family of amino acid transporters (BL00218, Fig. 5E-F). Among other interesting new relationships detected is one between giardins (GIA1_GIALA, GIA2_GIALA) and annexins (BL00223), both of which are cytoskeletal components, and another suggesting that mouse transplantation antigen (TUM8_MOUSE) is the first eukaryotic example of ribosomal protein L13 (BL00783, Fig. 5G).

Each of the 38 sequences reported in Table 2 was used to search the database of 803 patterns in PROSITE v. 10.0, the same database from which BLOCKS v. 6.0 was derived. Results were identical using either PATMAT (Wallace and Henikoff, 1992) or the MOTIFS program of the GCG package (Devereux *et al.,* 1984): in every case the sequence failed to detect the pattern or patterns representing the PROSITE group corresponding to that reported in Table 2.

Identification of repeated domains

Since each block typically represents a single protein motif, the presence of repeated motifs can be detected in a search as high scores for a single block at multiple positions within the query sequence. Among the set of protein queries tested, examples of repeated motifs were identified. In some cases, these multiple motifs were found in separate parts of the protein, such as for the previously undetected beta-transducin similarity found at several positions within PLAP_MOUSE (Fig. 6A). In other cases, a single block encompassed multiple copies of a repeat. This led to multiple high PATMAT scores for alignments that overlapped. An example is *E. coli* FirA, an uncatalogued member of the cysE/lacA/nodL acetyltransferase family (BL00101). Recently, Dicker and Seetharam detected an "isoleucine patch" within FirA and members of this family consisting of a 6 residue repeat present in FirA a total of 28 times (Dicker and Seetharam, 1992). The single 47 residue wide block representing this family in the Blocks Database includes 8 copies of this repeat. Multiple high PATMAT scores were reported for this block within FirA, including the top 3 scores in the search. In all, 18 high scores were reported, mostly at overlapping positions. In 9 cases, the successive high-scoring alignments were offset by 6 amino acids (Fig. 6B). Thus the 6-mer repeat within members of this family becomes obvious upon examination of search results.

Classification of suspected helicases

In some cases, families of related proteins are thought to belong to larger "superfamilies". However, the inclusion of a family into a superfamily is often difficult to determine by objective criteria (Henikoff, 1991). A case in point is a collection of distinct families that are claimed to belong to a superfamily of helicases. Several reports (Burgess *et al.,* 1990; Company *et al.,* 1991; Girdham and Glover, 1991; Davis *et al.,* 1992; Laurent *et al.,* 1992; Johnson *et al.,* 1992; Okabe *et al.,* 1992; Kuroda *et al.,* 1991) have suggested that two new families of proteins, those related to *S. cerevisiae* SNF2 and those related to *S. cerevisiae* PRP16 (the DE-H family), contain "helicase motifs". These motifs derive from alignments of likely RNA- and DNA-dependent helicases and other nucleic acid dependent ATPases (Gorbalenya *et al.,* 1989). However, helicase motifs are generally detected using manual procedures that lack negative controls inherent in computer-based database searches. Of specific concern is the fact that some of the motifs are common to many ATPases that are not involved in nucleic acid metabolism.

A Blocks Database (v. 5.0) was supplemented with blocks representing families used to derive the helicase motifs, and with blocks representing two new families in which these motifs have been reported. Each sequence from each family was used to search the database. Combined search results for all of the sequences from a single family are shown in Fig. 7 as ranges of expectant values for the detection of blocks representing each of the families. The diagonal values are for detection of a family to which a sequence is known to belong, and off-diagonal values represent potential cross-family relationships. For each of the two new families, SNF2 and DE-H, a single unequivocal cross-family relationship was detected. In contrast, no other cross-family relationships were reliably detected above background probabilities ($E > 10^{-3}$) in these searches, nor above what was seen for searches using control ABC family sequences which contain ATP-binding domains very similar to those found in the putative helicases.

A global relationship is found between the DE-H family and the family of cylindrical inclusion (CI) proteins from positive-stranded RNA viruses. This confirms the previous detection of this specific relationship by Koonin (1991). Fig. 8A shows the alignment of bovine diarrhea virus CI protein with sequence segments from blocks derived from CI proteins of other viruses, with 4 of 5 blocks detected ($E=1x10^{-8}$). Similarly, 4 of 5 DE-H family blocks were detected ($E=5x10^{-9}$) in the same search (Fig. 8B). In each case, the single best alignment of a segment within a block to the BVDV CI sequence differs from one block to another, with 4 different proteins represented. This illustrates an important advantage of searching a database of blocks over searching a databank of sequences: for different conserved segments, the closest similarities to a distant relative are distributed among different members of a family (see also Figs. 2 and 5 for examples).

The analysis also revealed an unequivocal global relationship between the SNF2 family and poxvirus DNA-dependent ATPases (the VATP family), as reported previously (Henikoff, 1993). This specific cross-family relationship was not reported by Bork and Koonin (1993) for the SNF2 family, nor by Koonin and Senkevich (1992) for the poxvirus proteins. Our analysis detected no other consistent relationships among hypothesized helicase superfamily members above the level seen for occasional background hits in these searches ($E=10^{-3}$). These frequent background hits might be attributable to well-known features that are common to ATP-binding proteins, such as are present in the ABC negative control group. We suggest that the delineation of specific cross-family relationships provides more useful information than is obtained from classification of

much more diverse sequences into a single superfamily (Gorbalenya *et al.,* 1989). It is worth noting that subsequent to submission of our initial report (Henikoff, 1993), SNF2_YEAST was revealed to be a DNA-stimulated ATPase without detectable helicase activity (Laurent *et al.,* 1993). Lack of helicase activity is a feature of well-studied members of the poxvirus DNA-dependent ATPases (*e. g.* Kunzi and Traktman, 1989).

DISCUSSION

We have described an approach to protein family classification that involves searching a database of protein blocks for both local and global similarities. In previous work, blocks representing the Tc1 family of transposase proteins were used to query a nucleotide sequence databank; new Tc1 family members were identified when a high scoring alignment of a block to a sequence entry was supported by other correctly-spaced block alignments (Henikoff, 1992). Evaluation was based on the rank of the supporting alignments in a search, leading to probability estimates of chance similarity that took advantage of the fact that each block was searched independently. Here we have reversed this basic approach to the more typical situation in which the query is a sequence of interest. The database that is searched includes families that are usually represented by more than one block; therefore independent detection of multiple blocks can be used to compute a global "expectant value". This value can be combined with an independent "anchor block score" for the best local alignment to arrive at an overall level of confidence.

For this searching approach to be most effective, scores should be based on biological realities as well as on a reasonable model of chance. To accomplish this goal, raw block alignment scores are normalized based on an empirical calibration procedure that involves using the block to query the full SWISS-PROT database. Normalization should compensate for any advantage that one block might have over another in a search. Indeed, distributions of anchor block frequencies appear to be well behaved using shuffled sequence queries, and do not reveal any subsets of blocks that stand out. Furthermore, expectant values reflect observed probabilities of multiple block hits both for shuffled sequence queries and for true negative sequence queries. These empirical results can justify an interpretation of these measures of similarity in terms of chance probabilities. So, an anchor score in the 99th percentile is one that is expected to occur among chance alignments once in 100 searches, and an expectant value of 10^{-6} is expected to occur by chance once in 1 million searches. Our discovery of interesting new relationships from a presumed true negative set of queries shows that these two measures can provide guidance for inferring homology.

We suggest that our empirical normalization procedure provides more realistic measures for determining biological significance than some others in common use. For example, BLAST computes a P-value based on a theoretical model that assumes protein sequences are random. However, Wootton and Federhen (1993) have found that about 40% of the proteins in SWISS-PROT do not conform to this model because of low complexity regions that frequently lead to inflated similarity scores. Not only are such regions typically omitted from blocks, but our empirical normalization procedure should prevent inflated scores.

Our approach also makes possible the ready detection of repeated motifs, seen as multiple high scoring regions of the query sequence for the same block. In the case of *E. coli* FirA (Fig. 6B), a repeated but diverged 6-mer repeat missed by several labs (Dicker and Seetharam, 1992) was easily detected as multiple overlapping high scores at 6 residue intervals.

An additional advantage of searching the Blocks Database for detection of homology is that a detected family relationship can be easily evaluated using the detailed documentation for PROSITE groups. For example, in their list of homologs to yeast chromosome III ORFs, Oliver *et al.* (1992) reported that YCR11c is a homolog of the *Drosophila white* gene. However, like the white protein, YCR11c is a member of the large ABC transport family; comparison to the blocks from this family shows YCR11c to be no more like the white protein than many other ABC proteins (unpublished results).

A different approach to protein classification is to search a database of simple patterns. The most comprehensive database of patterns is PROSITE (Bairoch, 1992), which includes one or more manually-derived pattern for each of the groups that we used to generate the Blocks Database. The popularity of searching simple patterns is evident from the number of programs available for searching PROSITE: 20 are listed in a recent compilation (Bairoch, 1992). However, of the 38 new sequence classifications reported here, not one of the corresponding patterns was detected by searching PROSITE. Furthermore, we are unaware of studies demonstrating that searching a database of simple patterns leads to the detection of relationships that are not easily detected using score-based methods such as that described here.

While our approach takes advantage of the many different groups already documented in PROSITE, it is not limited to those groups. For example, the "helicases" (Fig. 7) include families not represented in the catalog, or in the case of DE-H proteins, erroneously combined with the RAD3 proteins (Harosh and Deschavanne, 1991). Blocks were generated from these distinct families using the PROTOMAT system and these blocks were added to the Blocks Database. In this way, a suspected family of sequences is required to compete against known groups.

The Blocks Database is necessarily much less complete than the sequence databanks, since only catalogued groups with two or more members are represented. So at present, our approach can only supplement searches based on pairwise alignments (Pearson, 1990; Altschul *et al.,* 1990; Collins and Coulson, 1990). Nevertheless, searches of the Blocks Database provides a simplified and objective method for the detection and evaluation of distant family relationships, a challenging problem that has spawned numerous strategies. For example, Bork and associates (Bork *et al.,* 1992b; Bork *et al.,* 1992a) have described a combination of different approaches for evaluating distant relationships using ORFs from yeast chromosome III. While they were successful in detecting many previously undescribed relationships, they did not detect relationships involving two ORFs that were revealed in our tests. The typical biologist evaluating search data does not have the same level of sequence analysis expertise as Bork and associates. These authors have pointed to the need for new automated approaches to the problem (Bork *et al.,* 1992a). The automated searching system described here is one such approach. About 1400 people have used it as an electronic mail server designed for the the biologist with no special sequence analysis skills.

ACKNOWLEDGMENTS

This work was supported by a grant from NIH (RO1 GM29009). We thank Bill Engels and Mark Boguski for helpful comments. A portion of this work was performed at the 1992 "Recognizing Genes" workshop of the Aspen Center for Physics.

REFERENCES

Altschul, S. F., and Lipman, D. J. (1990). Protein database searches for multiple

alignments. *Proc. Natl. Acad. Sci. USA* **87:** 5509-5513.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215:** 403-410.

Bairoch, A. (1992). PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **20:** 2013-2018.

Bairoch, A., and Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **20:** 2019-2022.

Bork, P., and Koonin, E. V. (1993). An expanding family of helicases with the 'DEAD/H' superfamily. *Nucleic Acids Res.* **21:** 751-752.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. (1992a). Comprehensive sequence analysis of the 182 predicted open reading frames of yeast chromosome III. *Prot. Sci.* **1:** 1677-1690.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., and Sonnhammer, E. (1992b). What's in a genome? *Nature* **358:** 287.

Burgess, S., Couto, J. R., and Guthrie, C. (1990). A putative ATP binding protein influences the fidelity of branchpoint recognition in yeast splicing. *Cell* **60:** 705-717.

Cavener, D. R. (1992). GMC oxidoreductases: a newly defined family of proteins with diverse catalytic activities. *J. Mol. Biol.* **223:** 811-814.

Collins, J. F., and Coulson, A. F. W. (1990). Significance of protein sequence similarities. *Meth. Enzymol.* **183:** 474-487.

Company, M., Arenas, J., and Abelson, J. (1991). Requirement of the RNA helicase-like protein PRP22 for release from spliceosomes. *Nature* **349:** 487-493.

Davis, J. L., Kunisawa, R., and Thorner, J. (1992). A presumptive helicase (*MOTI* gene product) affects gene expression and is required for viability in the yeast *Saccaromyces cerevisiae*. *Mol. Cell. Biology* **12:** 1879-1892.

Devereux, J., Haeberli, P., and Smithies, O. (1984). A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* **12:** 387-395.

Dicker, I. B., and Seetharam, S. (1992). What is known about the structure and function of the *Escherichia coli* protein FirA? *Mol. Microbiol.* **6:** 817-823.

Fuchs, R. (1991). MacPattern: protein pattern searching on the Apple Macintosh. *CABIOS* **7:** 105-106.

Girdham, C. H., and Glover, D. M. (1991). Chromosome tangling and breakage at anaphase result from mutations in *lodestar,* a *Drosophila* gene encoding a putative

nucleoside triphosphate-binding protein. *Genes and Dev.* **5:** 1786-1799.

Gish, W., and States, D. J. (1993). Identification of protein coding regions by database similarity search. *Nature Gen.* **3:** 266-272.

Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P., and Blinov, V. M. (1989). Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. *Nucleic Acids Res.* **17:** 4713-4730.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* **84:** 4355-4358.

Harosh, I., and Deschavanne, P. (1991). The RAD3 gene is a member of the DEAH family RNA helicase-like protein. *Nucleic Acids Res.* **19:** 6331.

Harris, N., Hunter, L., and States, D. (1992). Megaclassification: discovering motifs in massive datastreams. *In* Proceedings of the National Conference on Artificial Intelligence. pp. 224-232. AAAI Press, Menlo Park.

Henikoff, S. (1991). Playing with blocks: Some pitfalls of forcing multiple alignments. *New Biol.* **3:** 1148-1154.

Henikoff, S. (1992). Detection of *Caenorhabditis* transposon homologs in diverse organisms. *New Biol.* **4:** 382-388.

Henikoff, S. (1993). Transcriptional activator components and poxvirus ATP-dependent ATPases comprise a single family. *Trends Biochem. Sci.* 18:291-292.

Henikoff, S., and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* **19:** 6565-6572.

Henikoff, S., Wallace, J. C., and Brown, J. P. (1990). Finding protein similarities with nucleotide sequence databases. *Meth. Enzymol.* **183:** 111-132.

Henikoff, S., Henikoff, J. G., Agus, S., and Wallace, J. C. (1993). Searching for homologies to protein blocks by electronic mail. *In* Automated DNA sequencing and analysis techniques. Academic Press, London.

Johnson, R. E., Henderson, S. T., Petes, T. D., Prakash, S., Bankmann, M., and Prakash, L. (1992). *Saccharomyces cerevisiae RAD5*-encoded DNA repair protein contains DNA helicase and zinc-binding sequence motifs and affects the stability of simple repetitive sequences in the genome. *Mol. Cell. Biology* **12:** 3807-3818.

Koonin (1991). Similarities in RNA helicases. *Nature* **352:** 290.

Koonin, E. V., and Senkevich, T. G. (1992). Vaccinia virus encodes four putative DNA and/or RNA helicases distanly related to each other. *J. Gen. Virology* **73:** 989-993.

Kunzi, M. S., and Traktman, P. (1989). Genetic Evidence for involvement of vaccinia virus DNA-dependent ATPase I in intermediate and late gene expression. *J. Virology* **63:** 3999-4010.

Kuroda, M. I., Kernan, M. J., Kreber, R., Ganetzky, B., and Baker, B. S. (1991). The maleless protein associates with the X chromosome to regulate dosage compensation in Drosophila. *Cell* **66:** 935-947.

Laurent, B. C., Yang, X., and Carlson, M. (1992). An essential *Saccaromyces cerevisiae* gene homologous to *SNF2* encodes a helicase-related protein in a new family. *Mol. Cell. Biology* **12:** 1893-1902.

Laurent, B. C., Treich, I., and Carlson, M. (1993). The yeast SNF2/SWI2 protein has DNA-stimulated ATPase activity required for transcriptional activation. *Genes and Dev.* **7:** 583-591.

Lindgren (1968). "Statistical Theory," Macmillan, New York.

Okabe, I., Bailey, L. C., Attree, O., Srinivasan, S., Perkel, J. M., Laurent, B. C., Carlson, M., Nelson, D. L., and Nussbaum, R. L. (1992). Cloning of human and bovine homologs of SNF2/SWI2: a global activator of transcription in yeast *S. cerevisiae*. *Nucleic Acids Res.* **20:** 4649-4655.

Oliver, S. G., van der Aart, Q. J. M., Agostoni-Carbone, M. L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., and Ballesta, J. P. G. (1992). The complete DNA sequence of yeast chromosome III. *Nature* **357:** 38-46.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.* **183:** 63-98.

Pongor, S., Skerl, V., Cserzo, M., Hatsagi, Z., Simon, G., and Bevilacqua, V. (1993). The SBASE protein domain library, release 2.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.* **21:** 3111-3115.

Smith, R. F., and Smith, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA* **87:** 118-122.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195-197.

Vrielink, A., Lloyd, L. F., and Blow, D. M. (1991). Crystal structure of cholesterol oxidase from *Brevibacterium sterolicum* refined at 1.8 A resolution. *J. Mol. Biol.* **219:** 533-554.

Wallace, J. C., and Henikoff, S. (1992). PATMAT: a searching and extraction program for sequence, pattern, and block queries and databases. *CABIOS* **8:** 249-254.

Wootton, J. C., and Federhen, S. (1993). Statistics of local complexity in amino acid

sequences and sequence databases. *Comput. Chem.* In press.

Table 1. Relationship between expectant value (E) and observed hit frequency

Expectant value interval within E interval	Observed frequency and probability of hits				
	a) Shuffled		?		b) Unshuffled
	Frequency	Observed P	?	Frequency	FrequencyObserved P
			?	(All)	(Pruned) (Pruned)
$E \leq 3.5 \times 10^{-6}$	0	0	?	62	0 0
$3.5 \times 10^{-6} < E \leq 3.5 \times 10^{-5}$	1	2.3×10^{-5}	?	11	0 0
$3.5 \times 10^{-5} < E \leq 3.5 \times 10^{-4}$	15	3.4×10^{-4}	?	63	18 3.9×10^{-4}
$3.5 \times 10^{-4} < E \leq 3.5 \times 10^{-3}$	124	2.8×10^{-3}	?	177	141 3.1×10^{-3}
$3.5 \times 10^{-3} < E \leq 3.5 \times 10^{-2}$	625	1.4×10^{-2}	?	774	740 1.6×10^{-2}
$3.5 \times 10^{-2} < E \leq 3.5 \times 10^{-1}$	3000	6.9×10^{-2}	?	3250	3227 7.1×10^{-2}
$3.5 \times 10^{-1} < E$	40018	9.1×10^{-1}	?	41685	41607 9.1×10^{-1}
Total hits	43783		?	46022	45733

Table 2. New classifications from searches of Blocks v. 6.0

BL#	Group description	Query ID	# Blocks	E	Score	Percentile[1]	Zone[2]
41	AraC bacterial regulators	MARA_ECOLI	1	1	1587	99.95	<1/1000
44	LysR bacterial regulators	DGDR_PSECE	1	1	1440	99.85	<1/100
59	Zinc alcohol dehydrogenases	MURA_ECOLI	4	6.8x10-11	1339	99.13	<1/7000
59		QOR_ECOLI	3	1.4x10-6	1694	100.00	<1/7000
59		VAT1_TORCA	3	1.5x10-7	1476	99.90	<1/7000
59		YCZ2_YEAST	3	1.4x10-6	1310	98.5	<1/7000
60	Iron alcohol dehydrogenases	TCBF_PSESP	3	1.0x10-7	1813	100.00	<1/7000
60		TFDF_ALCEU	4	1.1x10-8	1811	100.00	<1/7000
61	Short-chain alcohol dehydrogenases	CSGA_MYXXA	2	0.0056	1446	99.85	<1/7000
61		MAS1_AGRRA	3	3.3x10-5	1509	99.93	<1/7000
61		SPRE_RAT	3	1.3x10-6	1407	99.7	<1/7000
101	cysE/lacA/nodL acetyltransferases	LPXA_ECOLI	1	1	1499	99.93	<1/1000
195	Glutaredoxins	YRUB_CLOPA	2	0.016	1281	97.4	<1/1000
215	Energy transfer proteins	PMP4_CANB	3	0.0001	1384	99.5	<1/7000
216	Sugar transporters	TCR1_BACSU	3	0.0002	1274	97.0	<1/1000
218	Amino acid permeases	CADB_ECOLI	2	0.00023	1370	99.5	<1/7000
218		POTE_ECOLI	3	0.00026	1543	99.96	<1/7000
223	Annexins	GIA1_GIALA	3	0.00011	1574	99.96	<1/7000
223		GIA2_GIALA	2	0.0046	1578	99.96	<1/7000
275	Shiga/ricin toxins	JIP_HORVU	5	1.4x10-11	1275	97.0	<1/7000
282	Kazal serine protease inhibitors	FSA_PIG	1	1	1621	100.00	<1/7000
297	Heat shock protein 70	MREB_ECOLI	3	0.00017	1113	48.6	<1/100

462 Gamma-glutamyl transpeptidases PAC1_PSES3 7 5.1x10-20 1995
100.00<1/7000

489 Phage-type RNA polymerases YS21_MAIZE 5 2.5x10-11 1698
100.00<1/7000

491 Amino-P/proline peptidases AGS_AGRRA 3 3.7x10-5 1379
99.6 <1/7000

504 Fumarate reductase flavoproteins NADB_ECOLI 9 4.0x10-21 1960
100.00<1/7000

530 Ribonuclease T2s RNI_ECOLI 2 0.00029 1286
97.6 <1/1000

552 MerR bacterial regulators MERD_PSEAE 2 0.00042 1410
99.75 <1/7000

573 Class II flavoproteins CYMO_ACISP 3 0.0019 1257
95.8 <1/1000

623 GMC flavoproteins CHOD_BREST 3 5.3x10-5 1373
99.5 <1/7000

646 Ribosomal protein S13 YM08_PARTE 1 1 1623
100.00<1/7000

665 Dihydropicolinate synthetase NPL_ECOLI 4 5.4x10-11 1742
100.00<1/7000

678 Beta transducins PLAP_MOUSE 2 0.014 1428
99.83 <1/7000
678 YCD9_YEAST 2 0.00013 1516
99.94 <1/7000

703 Prokaryotic ornithine decarboxylase ADI_ECOLI 9 4.6x10-24 2324
100.00<1/7000

710 Phosphoglucomutases UREC_HELPY 3 7.8x10-7 1763
100.00<1/7000

758 ArgE/dapE/CPG2 peptidases CBPS_YEAST 3 4.3x10-5 1337
99.10 <1/7000

783 Ribosomal protein L13 TUM8_MOUSE 2 0.0002 1442
99.85 <1/7000

[1]Based on shuffled sequence searches where the maximum anchor block score was 1617 for 7,082 searches.
[2]Based on Fig. 4

FIG. 1. Overall strategy for searching a database of blocks. The PROTOMAT system is applied to a family of protein sequences, resulting in a "best path" of blocks, illustrated here for the example of Fig. 2. The Blocks Database consists of successive application of PROTOMAT to unique groups catalogued in PROSITE, including calibration of each block based on the results of searching SWISS-PROT (not shown). PATMAT converts each block to a search matrix and scores all possible alignments of the query with all blocks in the database, saving the top scoring alignments in rank order. BLOCKSORT starts with the top ranking alignment (block A from this best path) and determines whether the hit includes multiple blocks among the saved alignments that are correctly spaced along the query sequence. BLOCKSORT then computes an expectant value (E) to evaluate whether these supporting block alignments (B and D) are due to chance. The best path for the group is depicted with block widths (numbers below) and ranges of distances between blocks (numbers above). For the aligned segments shown in the bottom panel, the distance between each block is shown above.

FIG. 2. Example of BLOCKSORT output. This shows that YCZ2_YEAST is a member of the zinc-containing alcohol dehydrogenase family (BL00059 in Table 2). YCZ2_YEAST was compared with each block in the Blocks Database by PATMAT, which assigned each block alignment a score and ranked the scores. The BLOCKSORT program then collected all individually scored blocks for the BL00059 group as shown here. For each block alignment in the group, BLOCKSORT reports the rank, frame (always 1 for a protein query), score and location from the PATMAT results, where location refers to where the query aligns with the block. It also reports block strength (Henikoff and Henikoff, 1991) from the Blocks Database. Here, the A block is the anchor block because it ranks highest among all alignments with blocks in the BL00059 group. In this example, the BL00059A block also ranks highest in the search, but an anchor block need not have rank 1. The percentile for the anchor block score reported here (98.5) is based on the distribution of anchor block scores for shuffled sequence queries (Table 1a). The expectant value (E=1.4e-06) is computed for the B and D blocks in support of the A block. Below the expectant value, the four database blocks for the BL00059 group are mapped with the scale noted. The blocks are indicated by repeated upper case letters, and these are separated by minimum (:) and maximum (.) distances observed between blocks in known members of the family. The BL00059 blocks found in YCZ2_YEAST are mapped below the database blocks at the same scale for comparison. The first YCZ2_YEAST line includes the blocks in the hit (the anchor block A and the two supporting blocks B and D). The second YCZ2_YEAST line includes blocks listed for the group that are not included in the hit. The "<" before the A block on this line indicates that it aligns outside of the query map scale. While both the B and D blocks are correctly spaced from the anchor block and so are included as supporting blocks in the hit, the C block is too distant from the D block and is excluded. Only the higher ranked of the two alignments of YCZ2_YEAST with BL00059A is used. Below the map, an alignment is shown of the query sequence YCZ2_YEAST with the sequence closest to it in each database block included in the hit based on identical residues. For example, the BL00059A segment of YCZ2_YEAST aligns with the corresponding segment of ADHX_HORSE. In the alignments, distances between detected blocks are shown as (min, max): for the database sequences followed by the distance in the query sequence. So, in the Blocks Database distances between blocks BL00059A and BL00059B range from 10 to 14 residues, and the distance between these blocks in YCZ2_YEAST is 9 residues.

FIG. 3. A. Frequency distribution of rank 1 blocks for all 2,302 blocks in the Blocks Database (v. 6.0) resulting from searches using 7,082 shuffled query sequences (circles). Each point represents the number of searches in which different blocks ranked first. For example, 539 different blocks ranked first in two searches. This distribution is compared to the values expected for perfect Poisson frequencies (triangles). B. Frequency distribution of anchor blocks for all 43,783 hits reported in the 7,082 searches (solid line) compared to a normal distribution (dotted line). For example, 59 different blocks were anchor blocks (achieved a PATMAT score of at least 1000 and ranked highest for a group) in 10 searches.

FIG. 4. Occurrence of hits with respect to anchor score and expectant value (E) for shuffled queries. Points represent the best anchor block hit in all 7,082 searches for each expectant value interval (connected by solid lines), the 7th best anchor block hit (dotted line) and the 70th best anchor block hit (dashed line). Any combination of anchor score and E lying above the solid line is expected to occur by chance <1/7000 searches, and so forth. Percentiles are based on shuffled sequence anchor scores. Log E < -4 was never observed with shuffled sequence queries.

FIG. 5. Examples of new classifications found in searches using the real sequence test set. In each example, the alignment of the query segment with a sequence segment from blocks in the family is shown with the query segment on top and the block sequence segment below. The distance between successive query segments is shown in parentheses above the range of distances between block segments for all family members. Upper case indicates a match at that position between the query and any segment in the aligned block. Identities are boxed.

FIG. 6. Detection of repeats. See legend to Fig. 2. A) The A and B blocks of the beta-transducin family (BL00678) were detected at several positions within PLAP_MOUSE as shown on the query map. B) No map is provided for single block hits, however examination of the Location column shows that 18 copies of a 6-mer repeat were detected within the query sequence by the single block (see text).

FIG. 7. Ranges of expectant values (high/low) reported in searches of BLOCKS v. 5.0 (Henikoff and Henikoff, 1991) supplemented with the blocks representing several different families. Protein family names are displayed along the top with the number of blocks representing the family indicated in parentheses. Individual sequences from each family were used as queries, with the number of such sequences indicated in parentheses. Representative sequences (SWISS-PROT IDs) for each family are: POLG_BVDV (P80 sequence only) (CI), MLE_DROME (DE-H), NTP1_VACCV (VATP), SNF2 (SNF2_YEAST), DEAD (IF41_MOUSE), RAD3 (RAD3_YEAST), EX5B_ECOLI (RECB), and BROW_DROME (ABC). Expectant value ranges for the top background hit in each search are shown at the right.

FIG. 8. Alignments of segments from BVDV P80 protein with the most closely related segment from each block. Blocks are from A) positive-stranded RNA viral CI proteins and B) DE-H proteins. BVDV was excluded from the best path by the PROTOMAT system, which accounts for its absence from the blocks representing the viral CI proteins. See

legend to Fig. 5.

```
Block    Rank Frame Score Strength      Location Description
BL00059A   1     1   1310  2439         2-    42 Zinc-containing alcohol dehyd
BL00059A 371     1    825  2439         0-    40 Zinc-containing alcohol dehyd
BL00059B  15     1    984  1967        52-    77 Zinc-containing alcohol dehyd
BL00059C 105     1    891  2795        77-   134 Zinc-containing alcohol dehyd
BL00059D   2     1   1232  2388       174-   229 Zinc-containing alcohol dehyd

1310=98.5th percentile of anchor block scores for shuffled queries
E=1.4e-06 for BL00059D BL00059B in support of BL00059A
                      |-----  108 residues----|
   BL00059 AAAAAAAAA::.BBBBBB::.........CCCCCCCCCCCCC::::...DDDDDDDDDDDDD
YCZ2_YEAST AAAAAAAAA::BBBBBB:::::::::::::::::::::::DDDDDDDDDDDDD
YCZ2_YEAST <AAAAAAAA        CCCCCCCCCCCCC

BL00059A    <->A (1,35):1
ADHX_HORSE 9     AAVAWEAGKPVSIEEVEVAPPKAHEVRIKIIATAVCHTDAY
                  ||  | || |  | |          | ||  | |   ||
YCZ2_YEAST 2     KAVVIEdGKaVVkEgVPiPELeEGfVLIKtLAVAgnpTDwa

BL00059B   A<->B (10,14):9
ADH3_ASPNI 62    PLIGGHEGAGVVVAKGELVKDEDFKI
                   | |   ||  |  |  |   || |
YCZ2_YEAST 52    GsILGcdAAGqIVKLGPaVdpkDFsI

BL00059D   B<->D (78,122):96
 ADH_CLOBE 173   IGIGAVGLMGIAGAKLRGAGRIIGVGSRPICVEAAKFYGATDILNYKNGHIVDQVM
                  |  |||   |  |        | |          | |||     |     | |
YCZ2_YEAST 174   gGAtAVGqSLIQlAnKlnGftkIIVvAsrKhEKLlKEYGADqlfDYhDiDvVeQIk
```

A) PLAP_MOUSE vs. BL00678 (Beta-transducins)

```
Block    Rank Score Location
BL00678A   4  1125    69- 83
BL00678A   5  1076   109-123
BL00678A  13  1024   148-162
BL00678A  21   978   189-203
BL00678A  71   908   228-242
BL00678A  79   901    28- 42
BL00678B   1  1428    71- 82
BL00678B   2  1395   111-122
BL00678B  20   982    30- 41
BL00678B  47   936   191-202
BL00678B  52   928   150-161

1428=99.82th %-ile of anchor block scores for shuffled queries
E=0.014 for BL00678A in support of BL00678B

                         |-----  130 residues----|
   BL00678 AAA:::::.....................................................BB
PLAP_MOUSE                                                 :::::AAA:::::BB
PLAP_MOUSE               AAA     AAA     AAA     AAA     AAA
PLAP_MOUSE       BB              BB      BB     BB

BL00678A   <->A  (69,537):27       BL00678B  A<->B  (28,286):28
PR04_YEAST 364   VATGGGDGIINVWDI    CC4_YEAST 438   SGSTDRTVRVWD
                  ||||  |  |                       |||  | |   ||
PLAP_MOUSE 28    IATGGnDHNIcIfsL   PLAP_MOUSE 71    SGSWDtTaKVWl
```

B) FIRA_ECOLI vs BL00101 (cysE/lacA/nodL acetyltransferases)

```
Block    Rank Score Location
BL00101    1   1472  111-157
BL00101    2   1273  147-193
BL00101    3   1251  123-169
BL00101    8   1059   99-145
BL00101   12   1048  224-270
BL00101   11   1048  218-264
BL00101   20   1013   93-139
BL00101   23   1000  242-288
BL00101   33    985  141-187
BL00101   36    982  117-163
BL00101   43    974  260-306
BL00101   47    970  129-175
BL00101   61    951  202-248
BL00101  205    870  159-205
BL00101  235    862  261-307
BL00101  261    855  229-275
BL00101  269    853  196-242
BL00101  343    838  153-199

1472=99.89th %-ile of anchor block scores for shuffled queries
E=1.00 for BL00101

BL00101          (111,195):110
THGA_ECOLI 134   IGNNVWIGSHVVINPGVTIGDNSVIGAGSIVTKDIPPNVVAAGVPCR
```

```
                    |||| ||   ||  ||  |||  ||||  | |
FIRA_ECOLI 111   LGNNVsIGAnAVIesGVElGDNviIGAGcfVgKnskiGAgsrlwanv
```